Tutorial

# Probablity Distributions

29 SATURDAY
2022 JANUARY

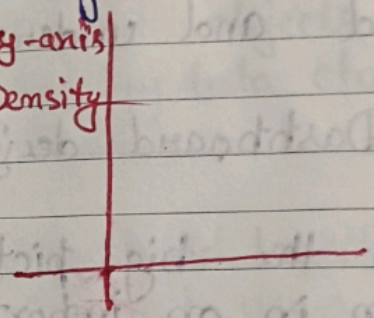
2021 DECEMBER

| M | T | W | T | F | S |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 29 | 30 | 31 | |

WEEK 05
029-336

## Probablity Distributions

A statistical model that shows possible outcomes of a particular event or course of action as well as the statistical likelihood of each event.




## How to go about this?

How do we use the collected business data (Sales volume, loan defaulters, Salary Hikes in an organization etc)?

1. The data values themselves are used directly in the simulation. This is called trace-driven simulation

30 SUNDAY

2. "Fit" a theoretical distribution to the data (and check whether that "fit" is good!)
3. The data values could be used to define an empirical distribution function in some way.

## What are these empirical distributions

- Using the data, we build our own distributions.
- How does one build a distribution?

- Essential building blocks:
Define the density/distribution functions.
Estimate the parameters (mean, standard deviation, etc).

## Empirical distributions

For ungrouped data:
Let $X_{(i)}$ denote the $i^{th}$ smallest of the $X_j$'s so that: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$.

$$F(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ \frac{i-1}{n-1} + \frac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})} & \text{if } X_{(i)} \leq x < X_{(i+1)} \text{ for } i = 1, 2, \dots n-1 \\ 1 & \text{if } X_{(n)} \leq x \end{cases}$$

For Grouped data:

$$G(x) = \begin{cases} 0 & \text{if } x < a_0 \\ G(a_{j-1}) + \frac{x - a_{j-1}}{a_j - a_{j-1}} \left[ G(a_j) - G(a_{j-1}) \right] & \text{if } a_{j-1} \leq x < a_j, j = 1, 2, \dots k \\ 1 & \text{if } a_k \leq x \end{cases}$$

FEBRUARY
MARCH
APRIL

1 TUESDAY
2022 FEBRUARY

2022 JANUARY

| S | M | T | W | T | F | S |
|---|---|---|---|---|---|---|
| 30 | 31 | | | | | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 23 | 24 | 25 | 26 | 27 | 28 | 29 |

WEEK 06
032-333

## The three approaches...

- Approach 1 is used to validate simulation model output for an existing system with the corresponding output for the system itself.
- Two drawback of approach 1: Simulation can only reproduce only what happened historically; and there is seldom enough data to make all simulation runs
- Approaches 2 and 3 avoid these shortcomings so that any value b/w minimum and maximum can be generated. So approaches 2 and 3 are preferred over approach 1.
- If theoretical distributions can be found that fits the observed data (approach 2), then it is preferred over approach 3

# Business Example

2022 JANUARY

| S | M | T | W | T | F | S |
|---|---|---|---|---|---|---|
| 30 | 31 | | | | | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 23 | 24 | 25 | 26 | 27 | 28 | 29 |

2 WEDNESDAY
2022 FEBRUARY

WEEK 06
033-332

~~Business Example~~

## Clues from summary statistics

- For the symmetric distribution mean and median should match. In the sample data, if these values are sufficiently closer to each other, we can think of a symmetric distribution (e.g. normal)
- Coefficient of variation (CV): (ratio of std. dev and mean $= \sigma/\mu$) for continous distributions. The CV=1 for exponential dist. If the histogram look like a slightly right skewed curve with CV>1, then lognormal could be better approximation of the distribution.
- Lexis ratio: same as CV for discrete distributions.
- Skewness(v): measure of symmetry of a distribution. For normal dist. v=0. For v>0, the distribution is skewed towards right (exponential dist, v=2). And for v<0, the distribution is skewed towards lyt. (left tail is longer than right tail)

## Parameter estimation

- One distribution is guessed, the next step is estimating the parameters of the distribution.
- Each distribution has a set of parameters.
  - ✓ Normal distribution has mean and standard deviation
  - ✓ Exponential distribution has a "$\lambda$".
- Most common method of parameter estimation: MLE (most likelihood estimation).

MARCH 2022

| M | T | W | T | F | S | S |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 28 | 29 | 30 | 31 | | | |

THURSDAY
FEBRUARY 2022
3

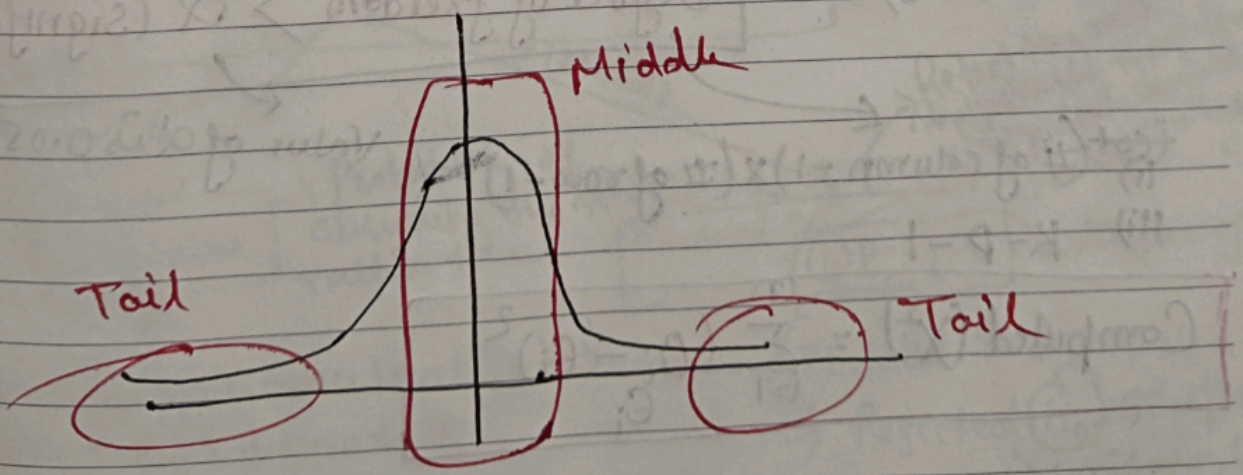
WEEK 06
034-331

## Goodness of fit

- For the input data, we have consumed a probability distribution.
- We also have estimated the parameters for the same.
- It can be checked by several methods:
1. Frequency comparison (a bit technical)
2. Probability plots (visual tool)
3. Goodness-of-fit tests (statistical test of goodness. Very widely used).

## Probability plots

- The Q-Q plot will amplify differences between the tails of the model distribution and the sample distribution.
- Whereas, the P-P plot will amplify the differences at the middle portion of the model and sample distribution.



4 FRIDAY
2022 FEBRUARY

2022 JANUARY

| S | M | T | W | T | F | S |
|---|---|---|---|---|---|---|
| 30 | 31 | | | | | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 23 | 24 | 25 | 26 | 27 | 28 | 29 |

WEEK 06
035-330

## Goodness-of-fit tests

- A goodness-of-fit test is a statistical hypothesis test that is used to asses formally whether the observations $X_1, X_2, X_3, \ldots, X_n$ are an independent sample from a particular distribution with function $\hat{F}$.

$H_0$: The $X_i$'s are IID (Independent Indentical Distributed) random variables with ~~dif~~ distribution function $\hat{F}$.

- Two famous tests:

1. Chi-Square test
2. Kolmogorov – Smirnov test

Tabulated $(\chi^2)$ = Degree of freedom > $\alpha$ (significance level)

→ Reject the Holmust hypothesis

Value of $\alpha$ is 0.05 or 0.01

(i) (# of column − 1) × (# of row − 1)

(ii) K − P − 1

$$\text{Computed } (\chi^2) = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

MARCH 2022

| M | T | W | T | F | S | S |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 28 | 29 | 30 | 31 | | | |

SATURDAY
FEBRUARY 2022
5

WEEK 06
036-329

# Tutorial — Overview of Chisquare GOF

## Chisquare GOF (Goodness of Fit Test)

This test is basically going to be used to see if the proposed distribution or hypothesized distribution for the population is a good fit for the population or not.

### Frequency Table

| Bins | Observed freq. | Expected freq. | $\chi^2$ |
|---|---|---|---|
| 1 | $O_1$ | $E_1$ | $(O_1-E_1)^2/E_1$ |
| 2 | $O_2$ | $E_2$ | $(O_2-E_2)^2/E_2$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| k | $O_k$ | $E_k$ | $(O_k-E_3)^2/E_k$ |

Test statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ (observed, expected)

Method 1: P-Value
- No. of parameters in population that you need to compute for given test
- P-value → Probability of observing the sample the null is true

P-value ≤ Significance level ($\alpha$) ⇒ Reject $H_0$ (dependent)

$H_a$ → The prob. is not uniform (not-independent)

Method 2: Tabulated $\chi^2$
- Reject the null hypothesis that the categorical variables are independent
- no. of bins
- $(k-1), \alpha$ → Degree of freedom

$\chi^2_{Tabulated} < \chi^2_{computed}$

⇒ Rejected $H_0$ → the population is uniform (independent)

Population Follows a

K-2-1 (i) Uniform Distribution ⇒ two parameters will be calculating ⇒ lower value (a), upper value (b) ⇒ So for uniform distribution P = 2 (a, b)

K-1-1 (ii) Poisson distribution ⇒ only one parameter (λ) ⇒ So P = 1 (λ)

K-2-1 (iii) Normal Distribution ⇒ two parameters ~~(μ, k)~~ $(\mu, \sigma^2)$ ⇒ Two parameters P = 2 $(\mu, \sigma^2)$

# Tutorial on performing Chisquare GOF for a Uniform Distribution

MARCH 2022

| M | T | W | T | F | S | S |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 28 | 29 | 30 | 31 | | | |

TUESDAY
FEBRUARY 2022
8

WEEK 07

Step1:- Descriptive Statistics
Step2:- Visual tools
} Narrowing down on possible guesses for the parent distribution

Step3:- Start creating the frequency table (computing the test statistics)

Step4:- Conclusion

($H_0$) Null Hypothesis:- The given data follows Uniform distribution.

($H_a$) Alternate Hypothesis:- The given data does not follow Uniform Distribution.

MARCH

APRIL